Hierarchical Models for RNA-seq

Michael Love

Hierarchical Models for RNA-seq

Michael Love
Dept of Biostatistics
Dept of Genetics

DNA => RNA

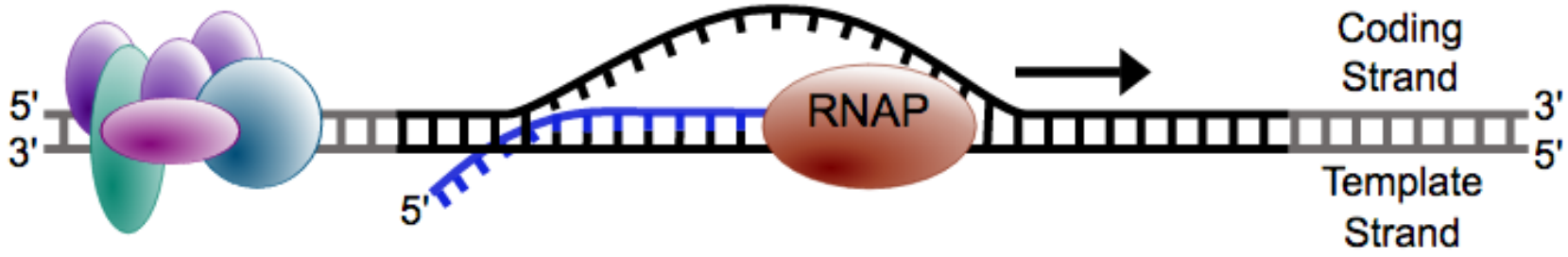


Figure 1:

Figure 2:

Why measure RNA: molecular phenotype

Figure 3:

Why measure RNA: tissue diversity

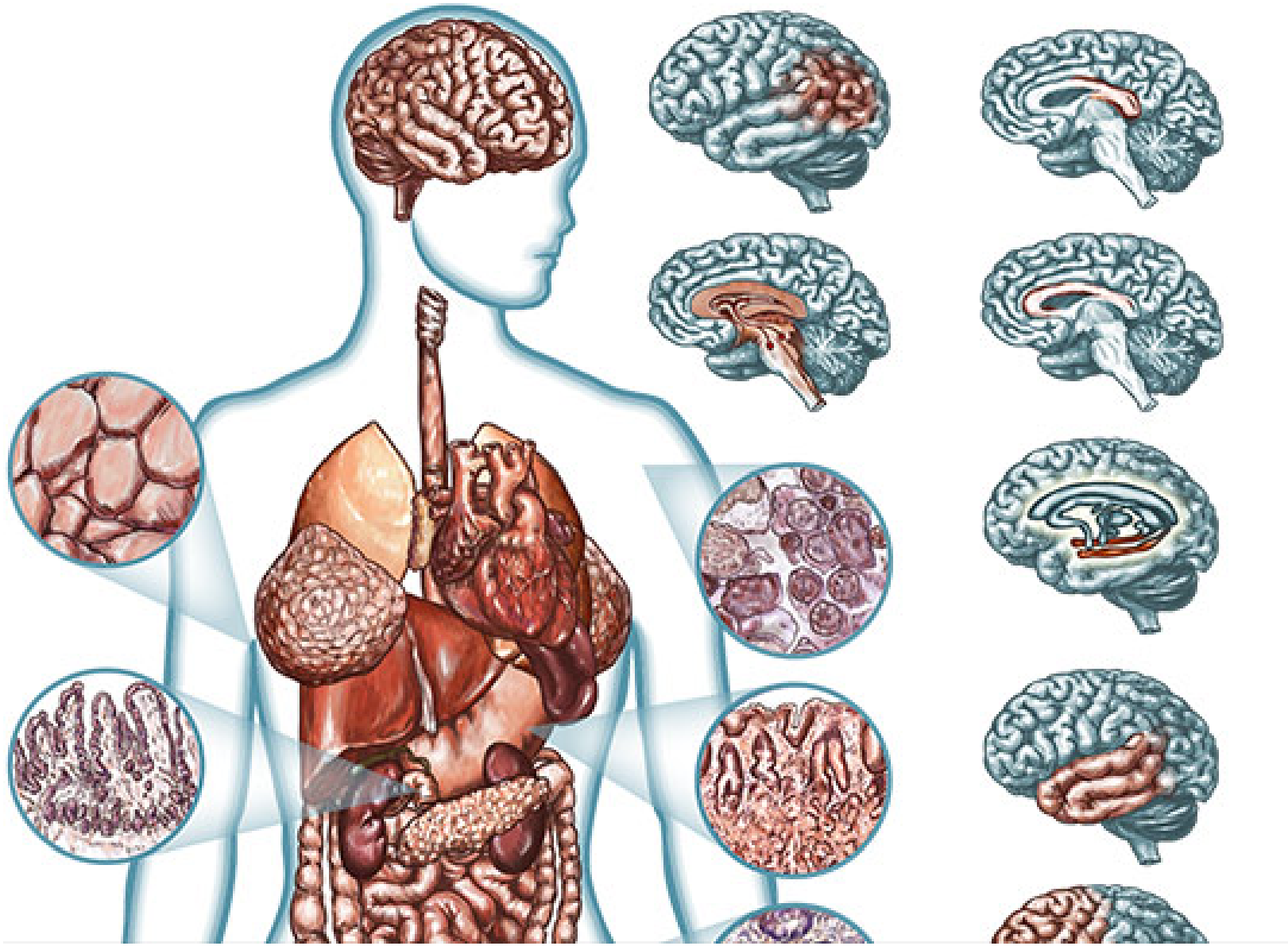


Figure 4:

Why measure RNA: tissue diversity

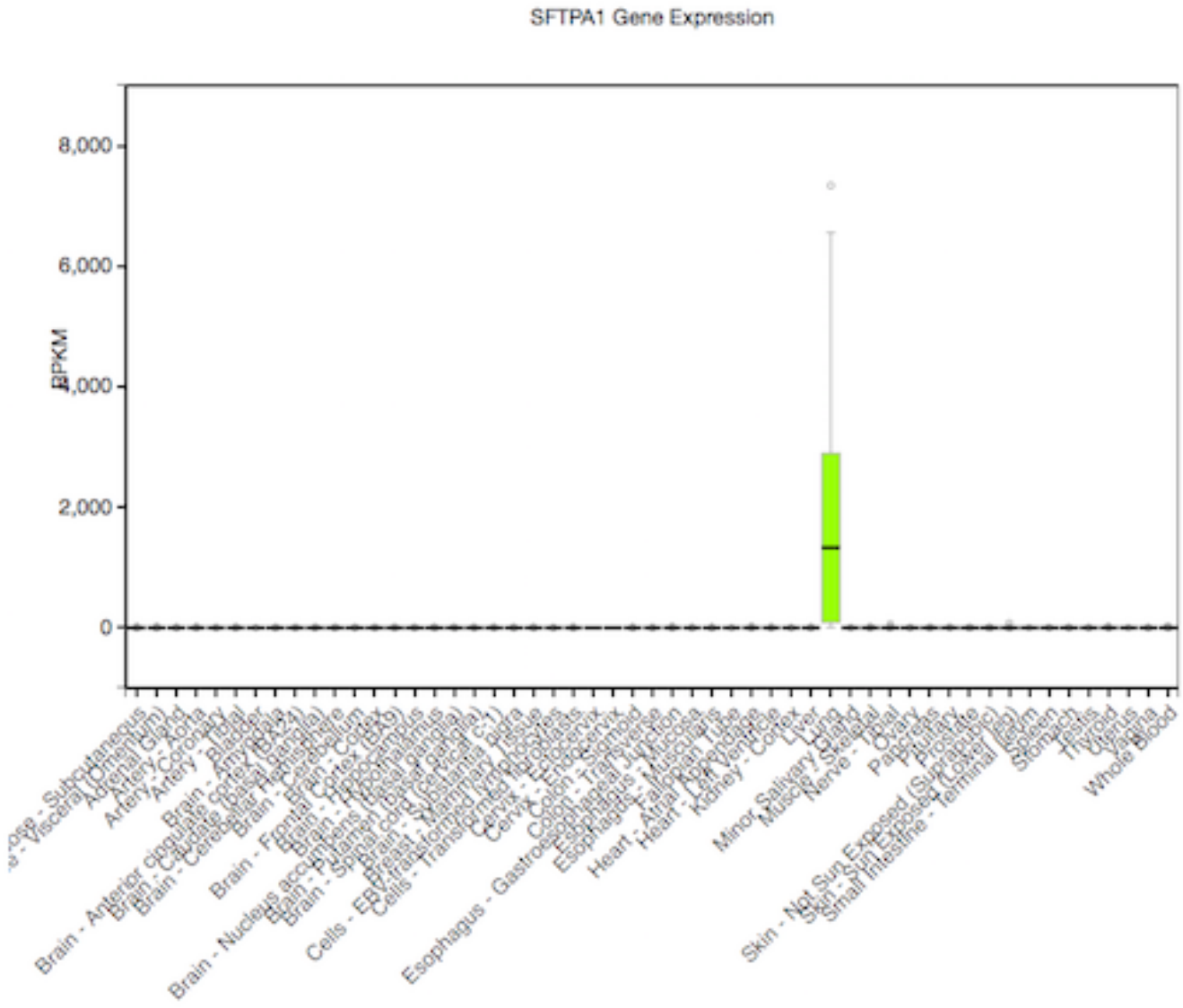


Figure 5:

Why measure RNA: within tissue over time

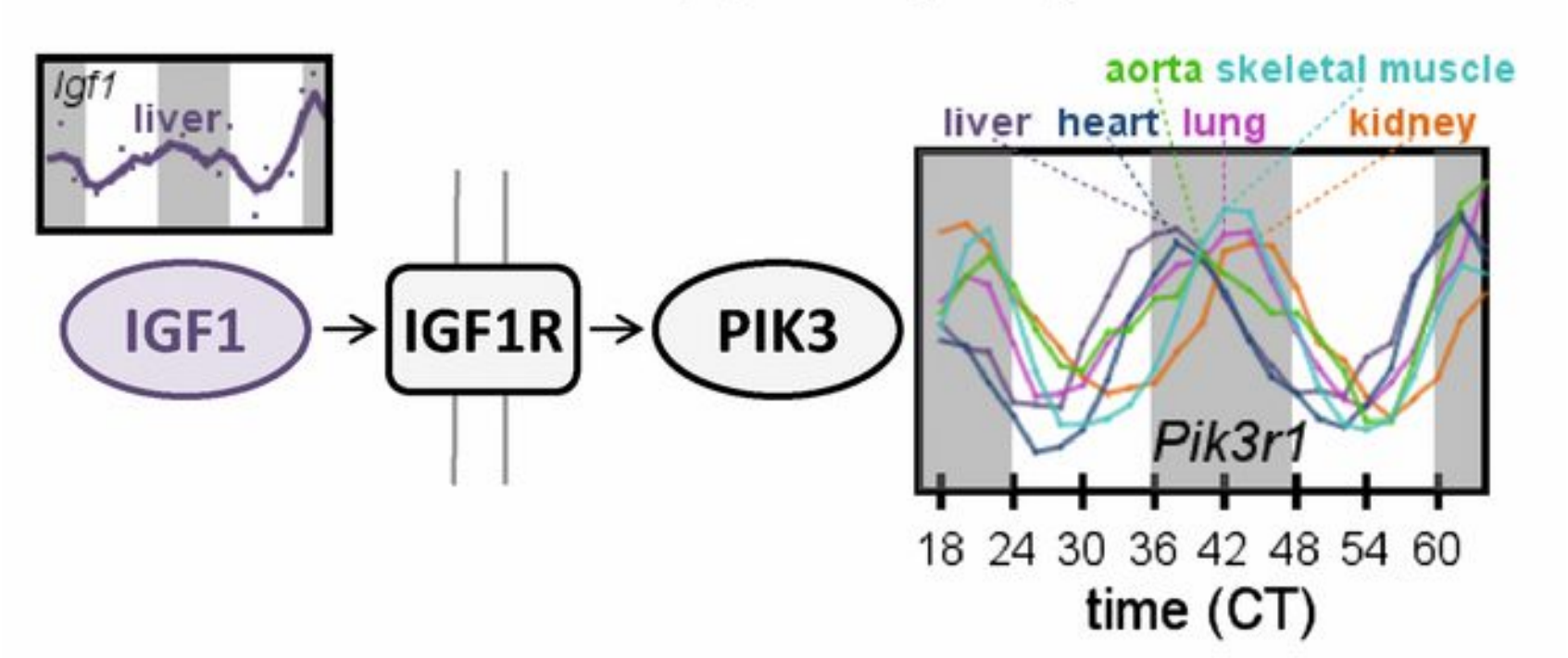


Figure 6:

Zhang, et al. Circadian gene expression atlas (2014)

Why measure RNA: disease sub-types

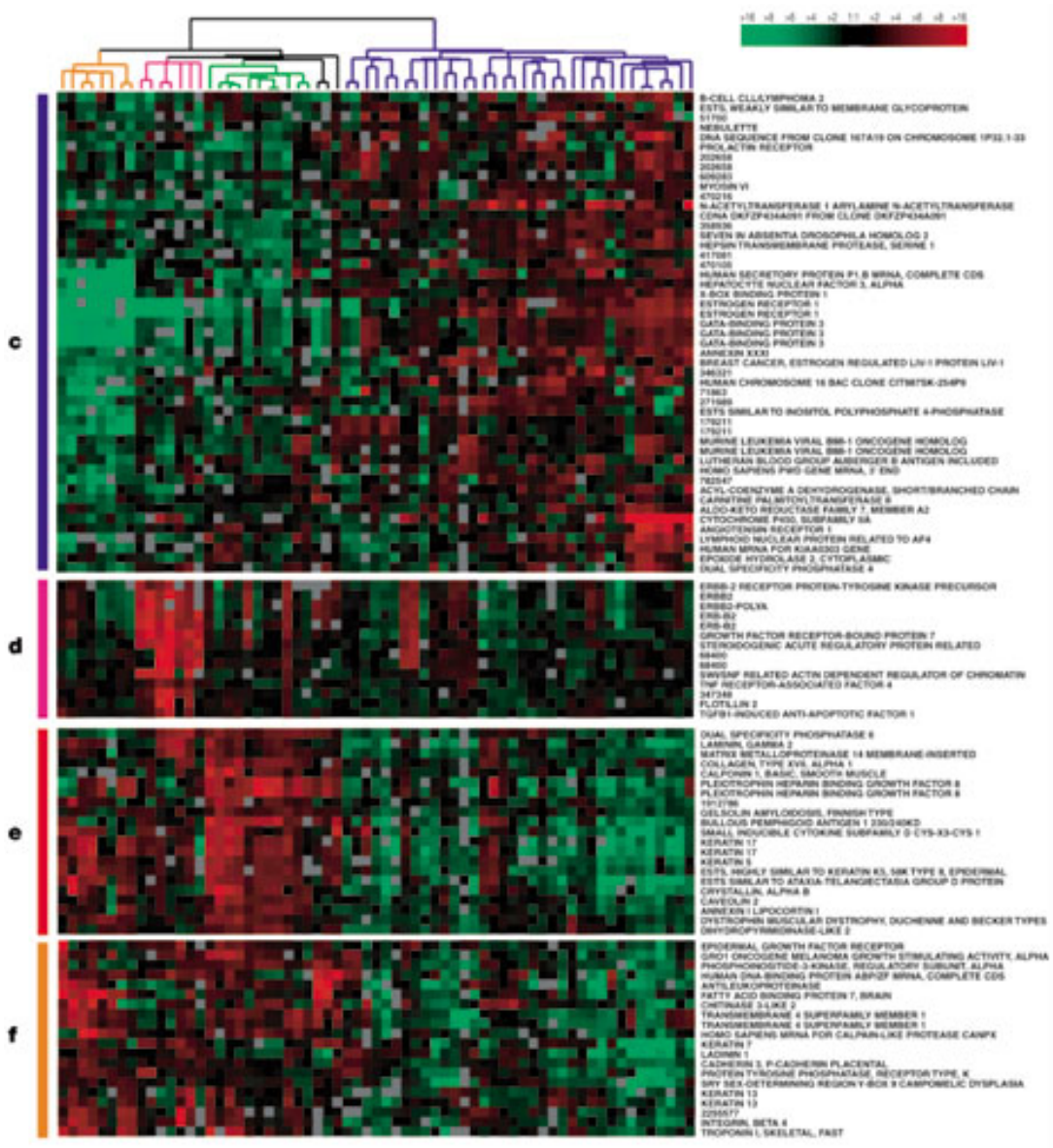


Figure 7:

Step back: pre-sequencing

Before sequencing was microarray

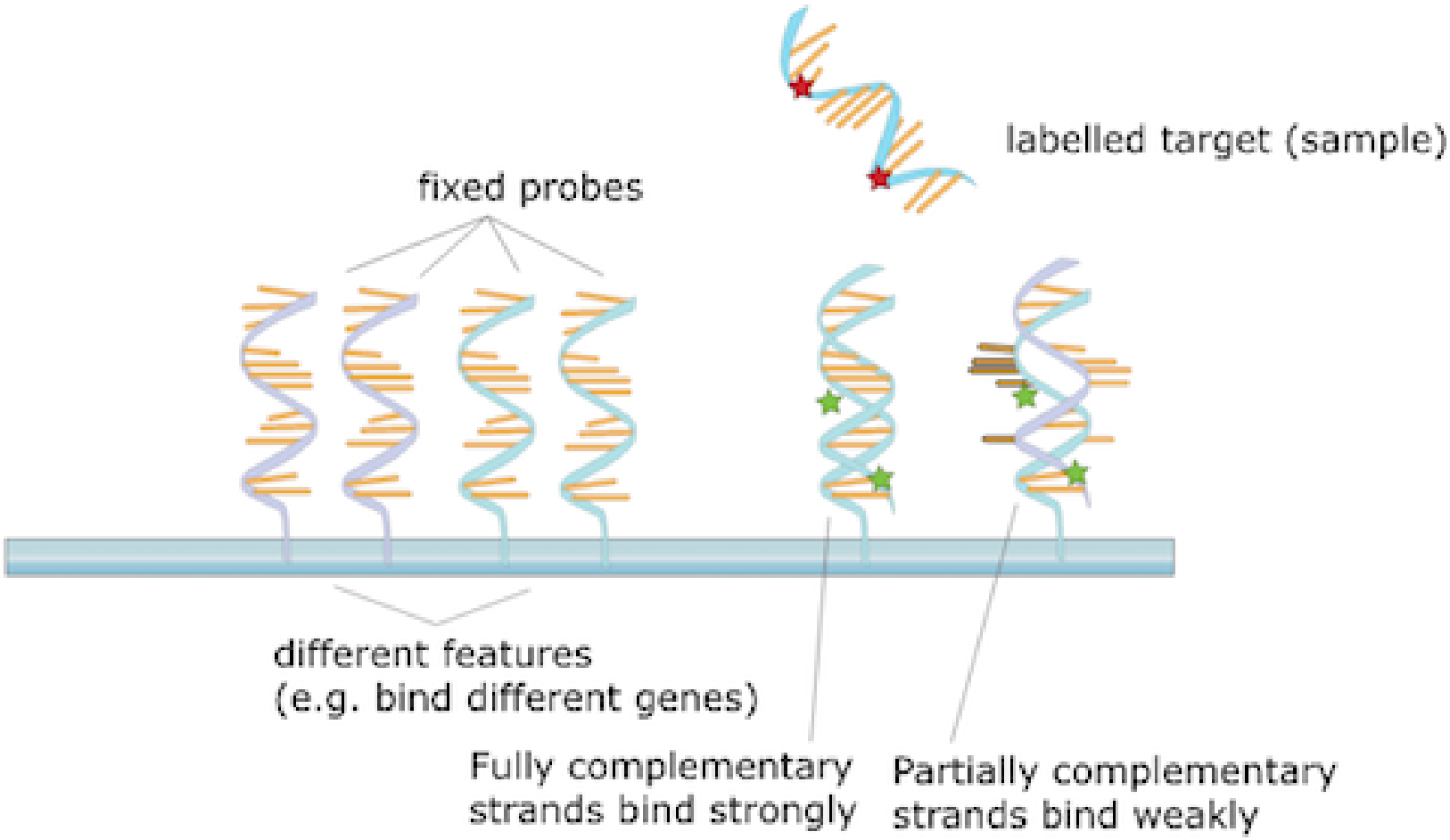
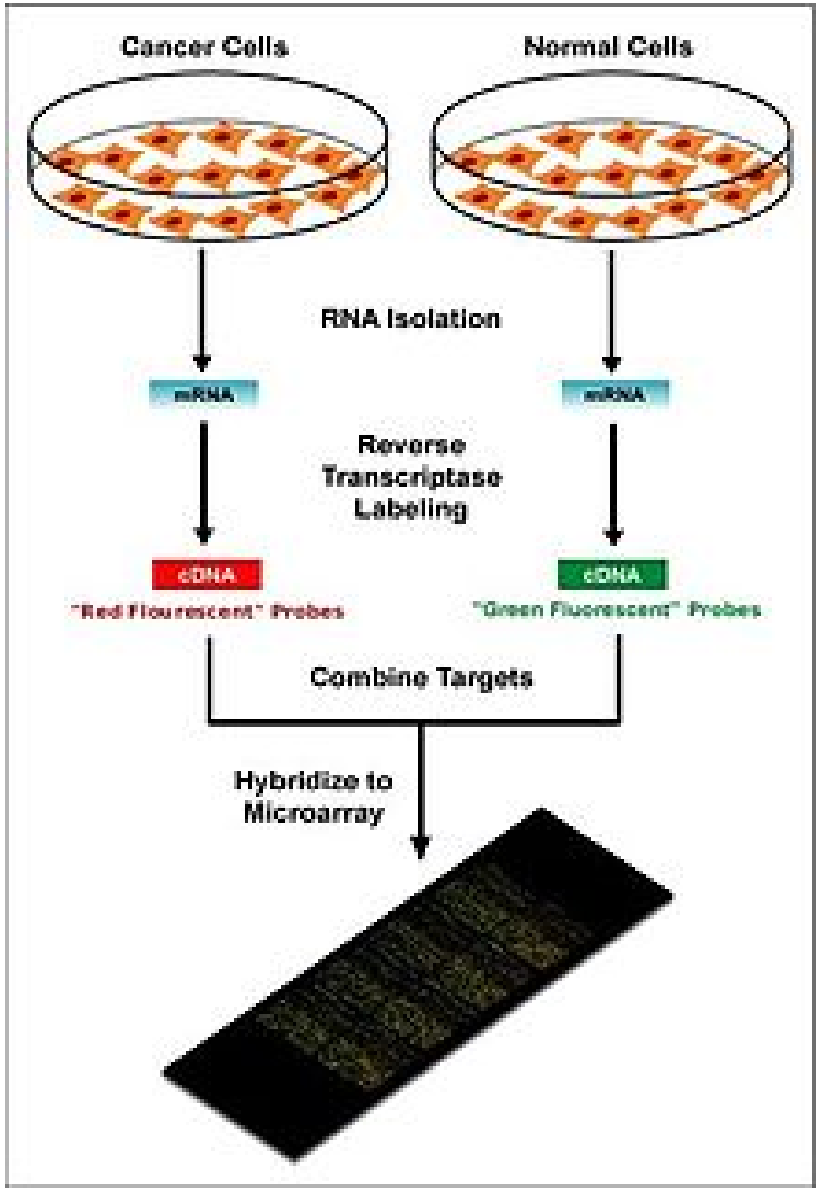


Figure 8:

Step back: pre-sequencing

Signal was captured light (positive, "continuous")



Motivating problem

- Gene expression for i=1,. . . ,N genes and j=1,. . . ,M samples
- log of gene expression values are in a tall matrix X
- log here is convenient because gene expression is non-negative and has a long tail
- 2 equal sized groups of samples A and B

$$X_{ij} \sim N(\mu_{ij}, \sigma_i)$$
$$\mu_{ij} = \mu_{i0}, \quad j \in A$$
$$\mu_{ij} = \mu_{i0} + \delta_i, \quad j \in B$$

$\delta_i \neq 0$ implies DE (differential expression)

Note σ_i

This is *critical*: different genes i have different amount of variability.

$$X_{ij} \sim N(\mu_{ij}, \sigma_i)$$
$$\mu_{ij} = \mu_{i0}, \quad j \in A$$
$$\mu_{ij} = \mu_{i0} + \delta_i, \quad j \in B$$

Goal of differential expression testing

- Find a set of genes for which $\delta_i \neq 0$
- And which obeys false discovery rate bounds
- For genes in our set G at FDR threshold z

$$E\left(\sum_{i \in G} 1_{\{\delta_i = 0\}}\right) \leq |G|\, z$$

Is this realistic?

- Can we accomplish this if all $\delta_i \neq 0$
- no, because methods often rely on global scaling normalization
- Are any $\delta_i = 0$?
- maybe not, but many are very small for controlled experiment

Is this realistic?

- What about σ_i for both groups?
- often this is enough, larger variance dominates
- not for single cell experiments
- More complex parametric models: baySeq
- Non-parametric: SAM / SAMseq

Back to the model

$$X_{ij} \sim N(\mu_{ij}, \sigma_i)$$
$$\mu_{ij} = \mu_{i0}, \quad j \in A$$
$$\mu_{ij} = \mu_{i0} + \delta_i, \quad j \in B$$

- N $= 5000$, M $= 6$
- $\delta_i = 0$ for 90%
- $\delta_i = \pm 2$ for 10%
- $\sigma_i \sim \Gamma(10, 10)$

Distribution of σ_i

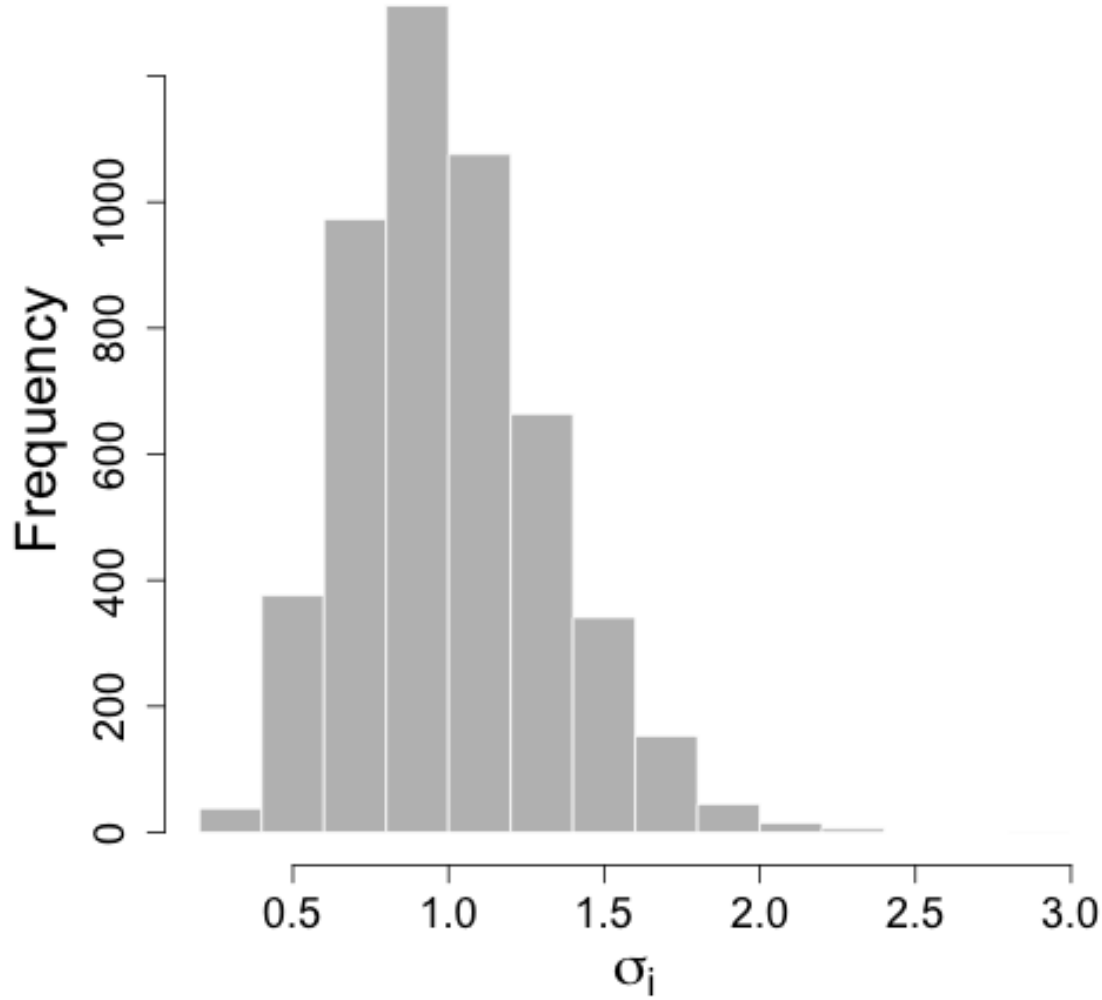


Figure 10: plot of chunk sigmadist

Try simple row t-tests

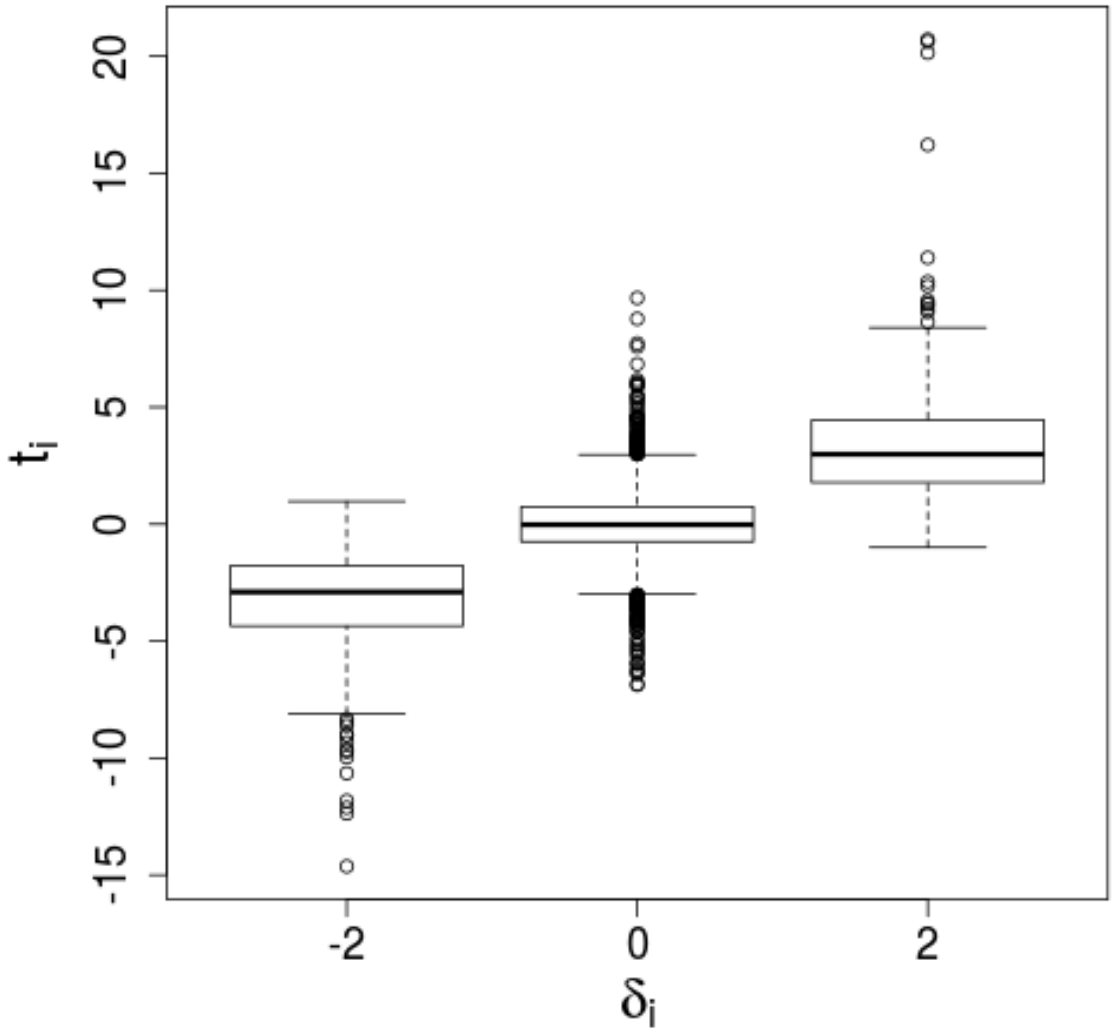


Figure 11: plot of chunk boxt

Just looking at ranks

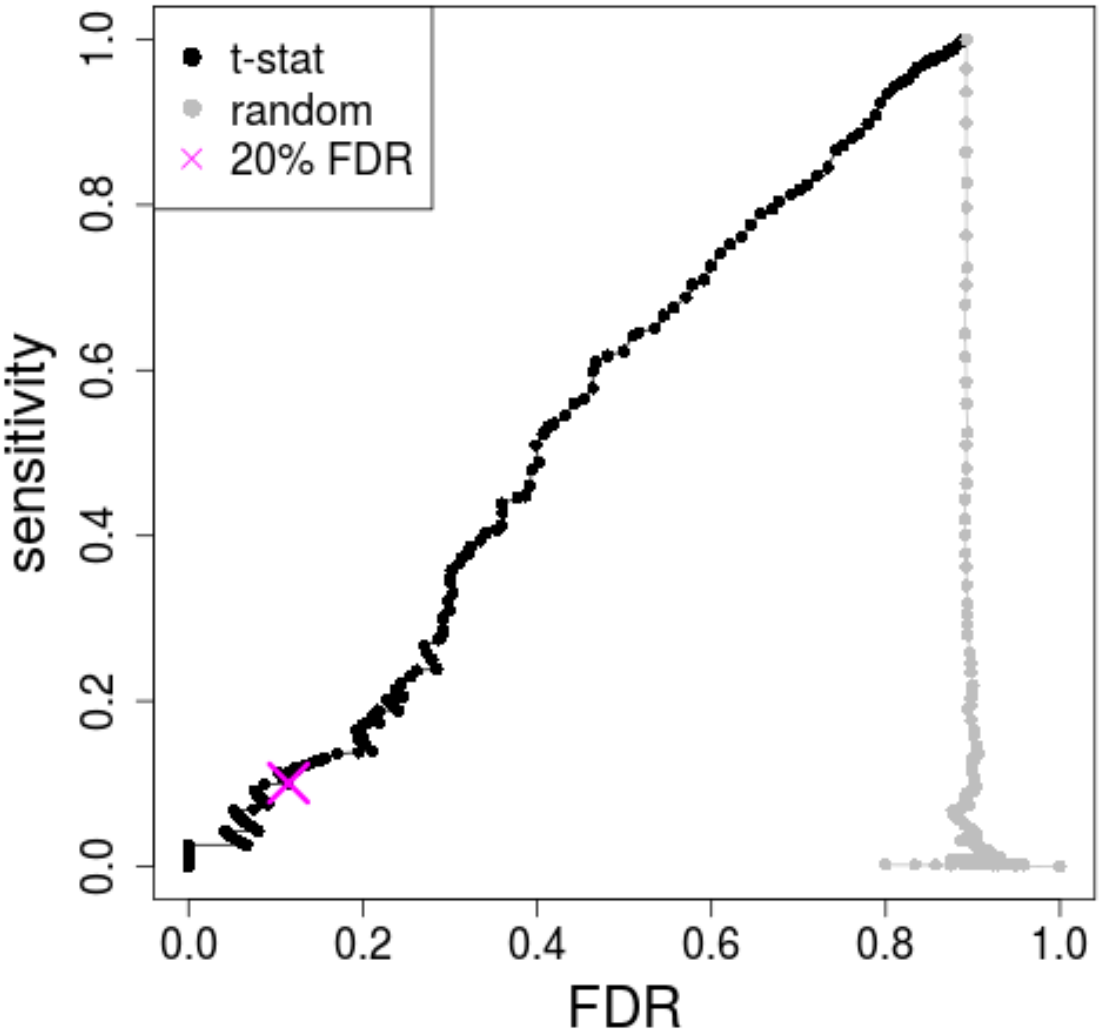
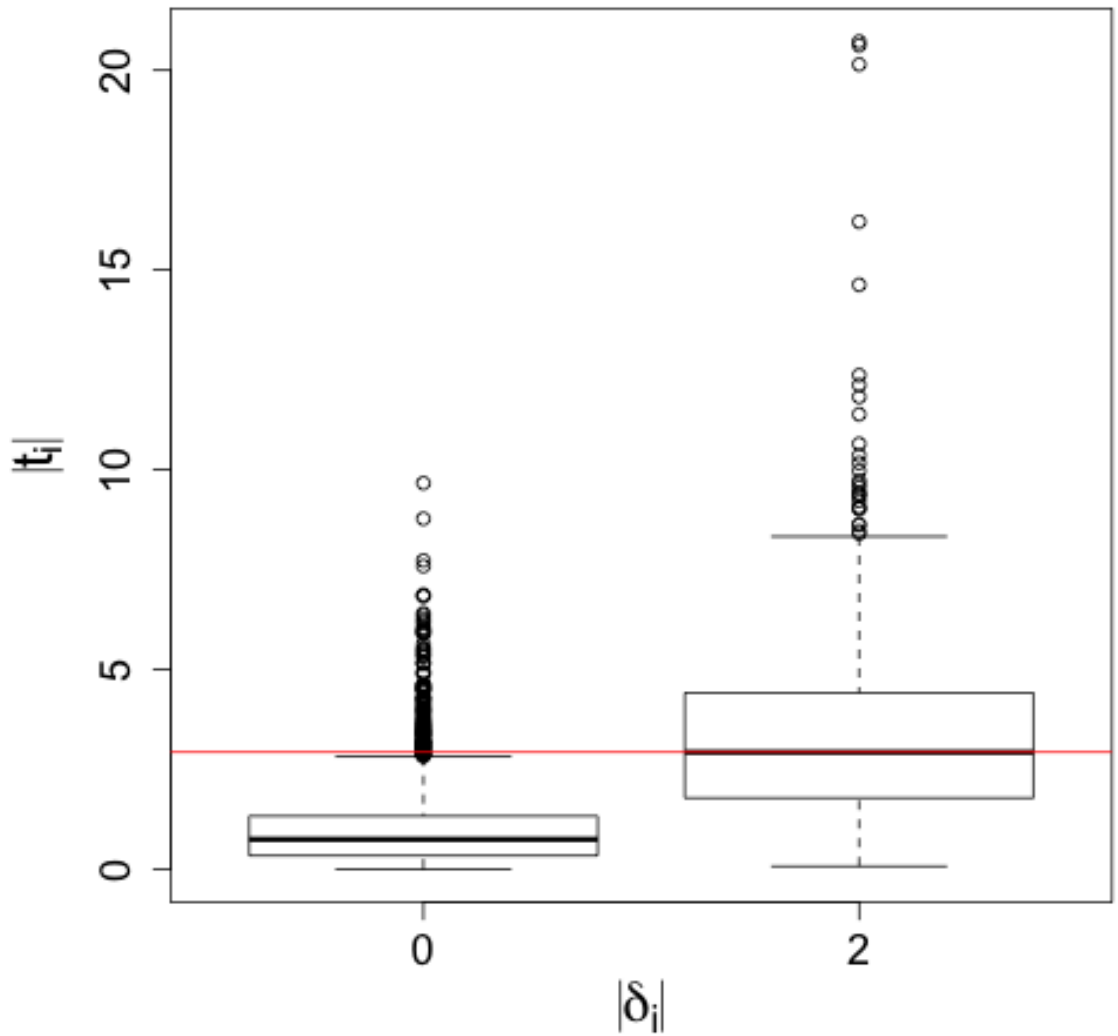


Figure 12: plot of chunk roc

Characterize the false positives

$$\mathrm{med}(t) \equiv \mathrm{median}(|t_i|) \text{ for } i : \delta_i \neq 0$$



Estimates of σ_i

$$\text{med}(t) \equiv \text{median}(|t_i|) \text{ for } i : \delta_i \neq 0$$

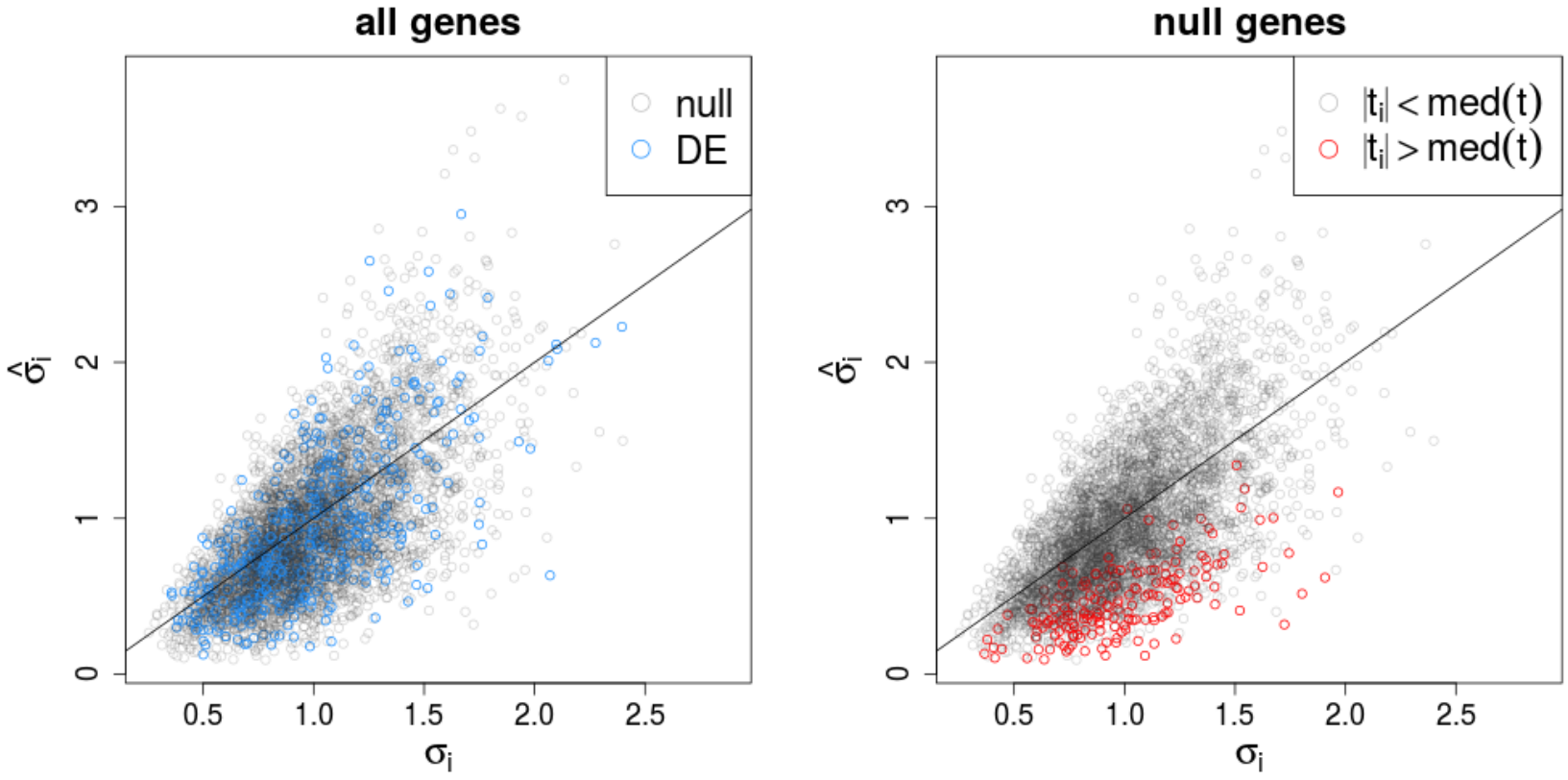


Figure 14: plot of chunk fp

New estimator for σ_i

$$\bar{\sigma} = \frac{1}{N}\sum_{i=1}^{N} \hat{\sigma}_i$$

$$\tilde{\sigma}_i^B = B\bar{\sigma} + (1 - B)\hat{\sigma}_i$$

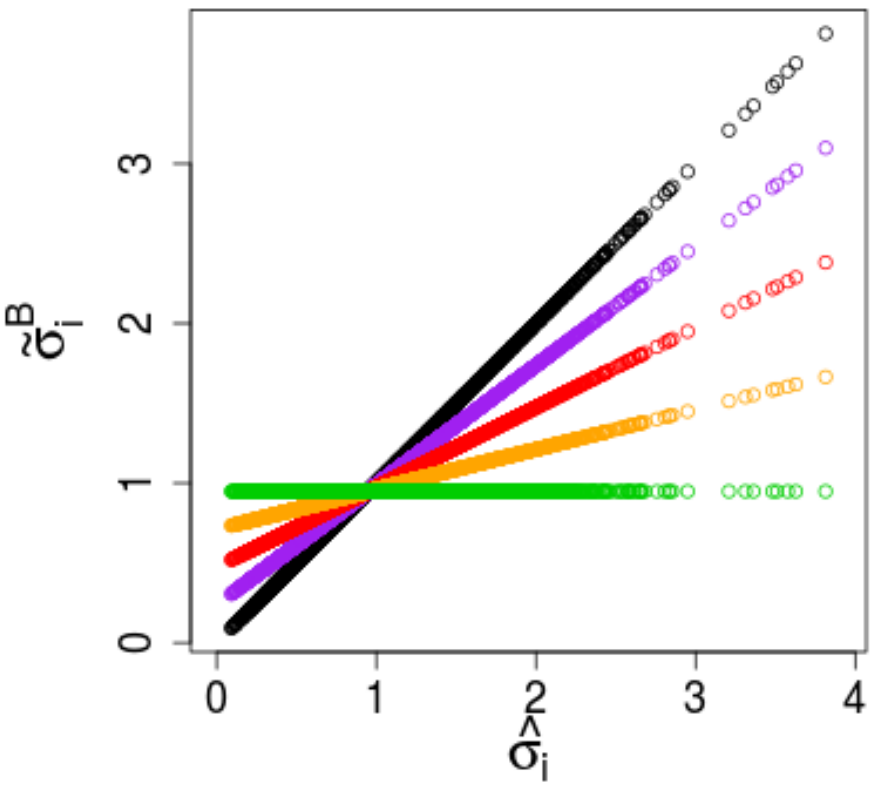


Figure 15: plot of chunk tildesigma

New estimator performance by rank

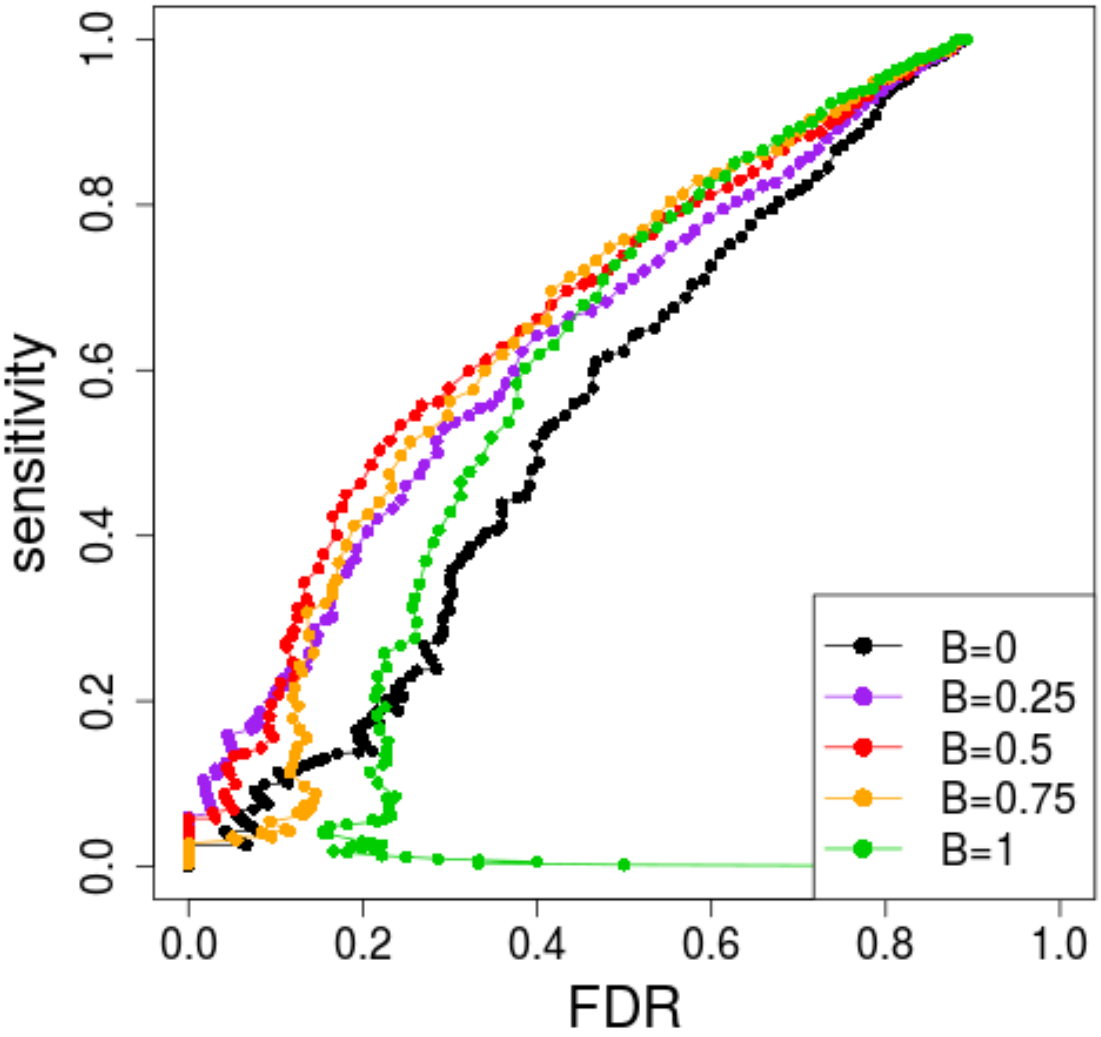


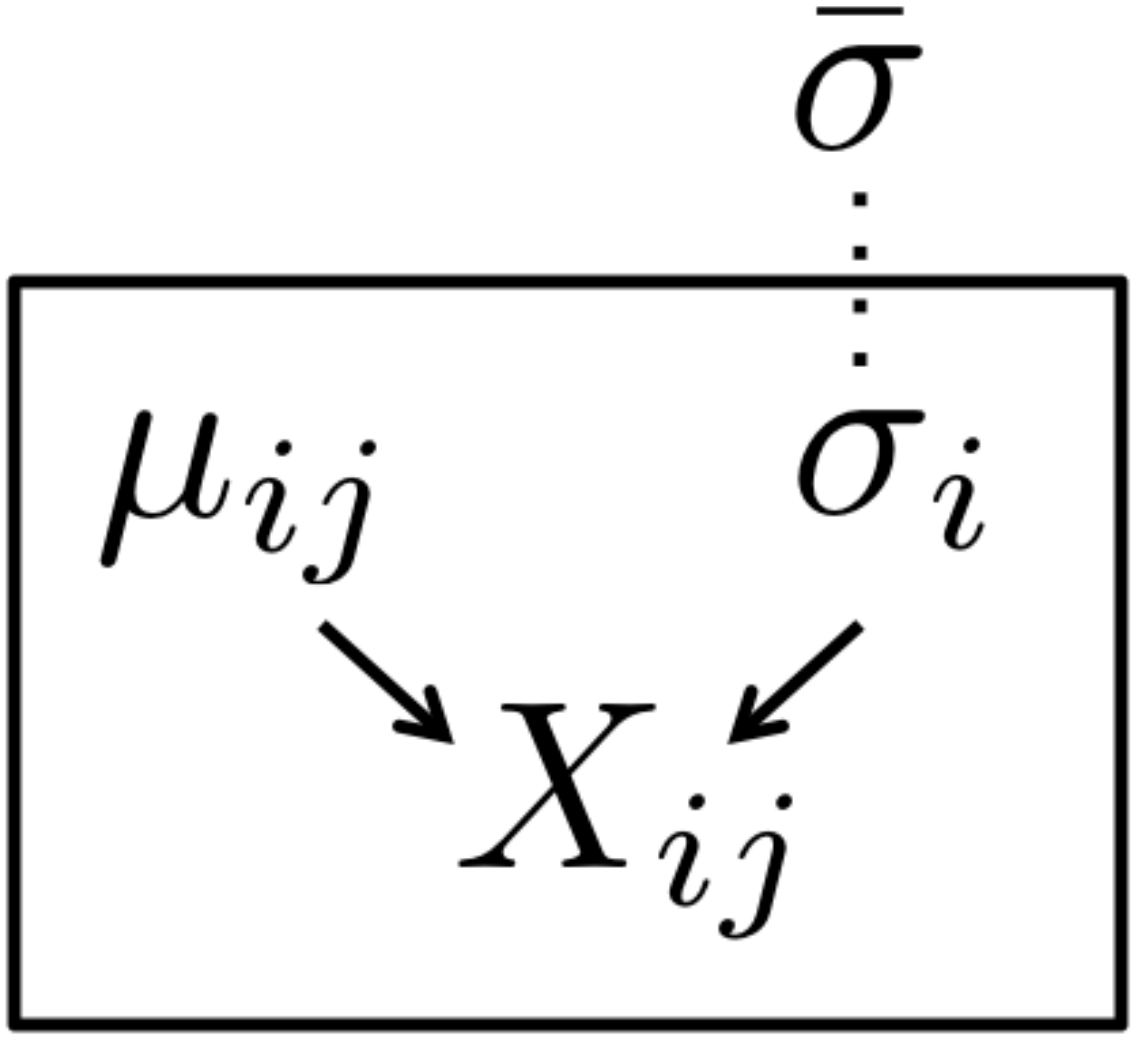
Figure 16: plot of chunk roc2

Summary

- Top false positives were coming from genes with too low $\hat{\sigma}_i$
- Replace $\hat{\sigma}_i$ with an estimate which is closer to $\bar{\sigma}$
- Depending on "close", new estimator dominates at all thresholds

How is this hierarchical?

Not your standard diagram, need to formalize



limma

- Smyth, G. K. (2004) Linear models and empirical Bayes methods for assessing differential expression in microarray experiments
- Developed the hierarchical model introduced by Lonnstedt and Speed (2002) for single sample into method for any experiment represented as linear model

$$\frac{1}{\sigma_i^2} \sim \frac{1}{d_0 \sigma_0^2} \chi^2_{d0}$$

Why inverse χ^2?

- Conjugacy provides closed form solution
- Posterior mean for $1/\sigma_i^2$ given $\hat{\sigma}_i^2$ is $1/\tilde{\sigma}_i^2$ with

$$\tilde{\sigma}_i^2 = \frac{d_0\hat{\sigma}_0^2 + d_i\hat{\sigma}_i^2}{d_0 + d_i}$$

And d_i as the standard residual degrees of freedom

Note that d_0 controls B

$$
\begin{aligned}
\tilde{\sigma}_i^2 &= \frac{d_0\hat{\sigma}_0^2 + d_i\hat{\sigma}_i^2}{d_0 + d_i} \\
&= \left(\frac{d_0}{d_0 + d_i}\right)\hat{\sigma}_0^2 + \left(\frac{d_i}{d_0 + d_i}\right)\hat{\sigma}_i^2 \\
&= B\hat{\sigma}_0^2 + (1 - B)\hat{\sigma}_i^2
\end{aligned}
$$

Proper hierarchical model

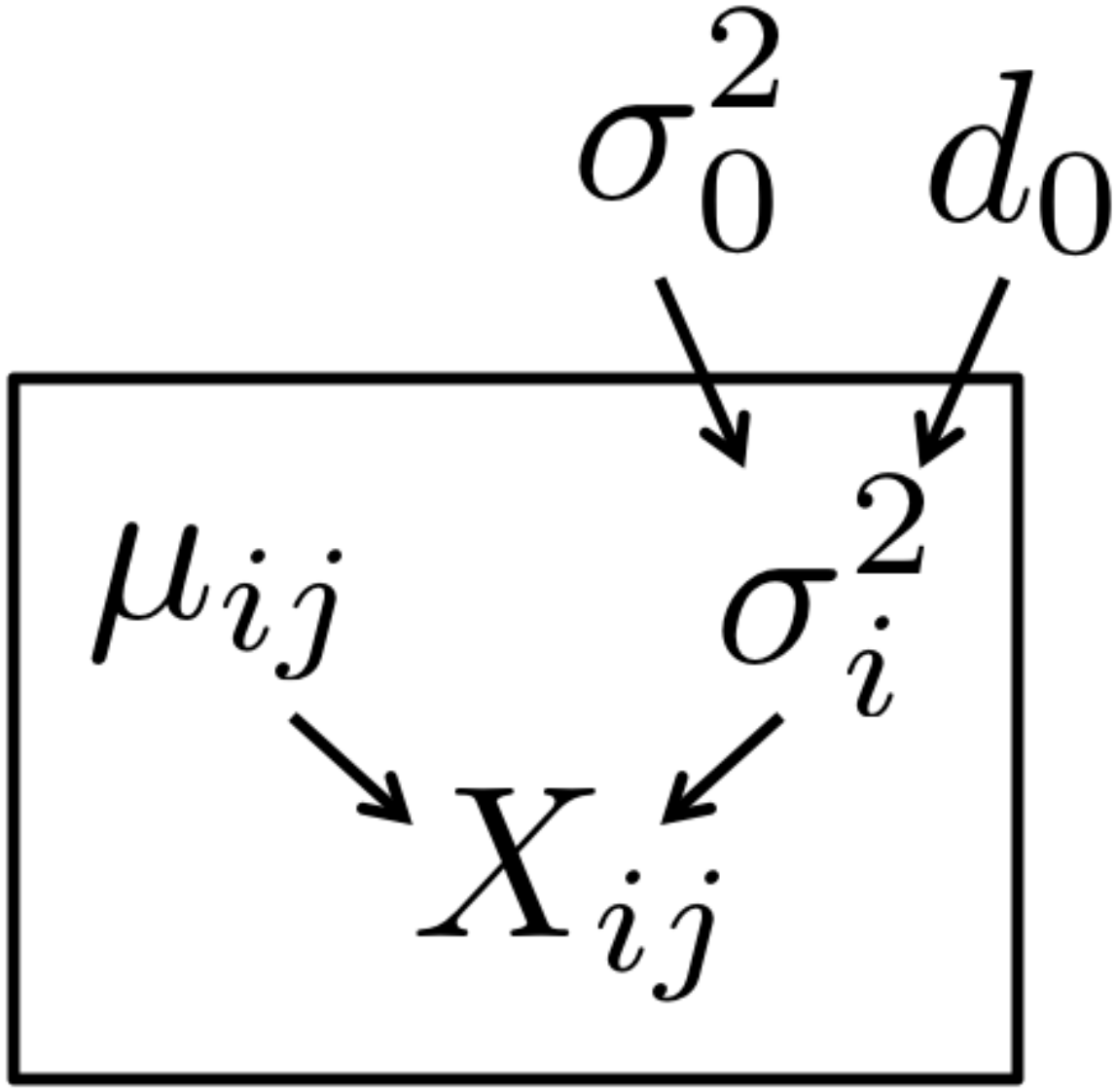


Figure 18:

Estimation of hyperparameters

- Need to estimate $d_0, \hat{\sigma}_0^2$, which control strength and location of *shrinkage* or *moderation*
- $d_0, \hat{\sigma}_0^2$ estimated via first two moments of $\log \hat{\sigma}_i^2$
- (Also need to estimate υ_0, another parameter giving variance of coefficients)

limma vs. naive estimators by rank

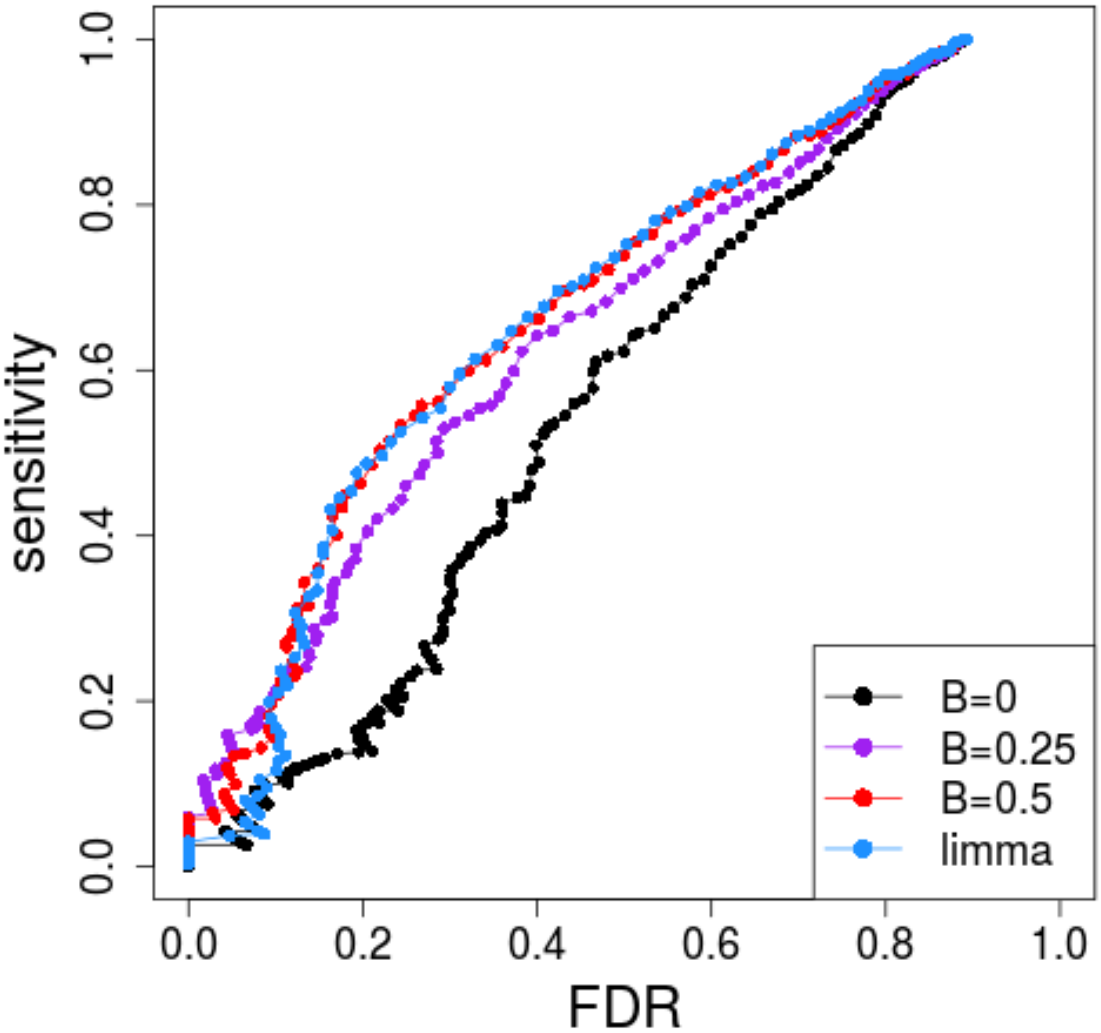


Figure 19: plot of chunk roc3

Rank is not the full picture

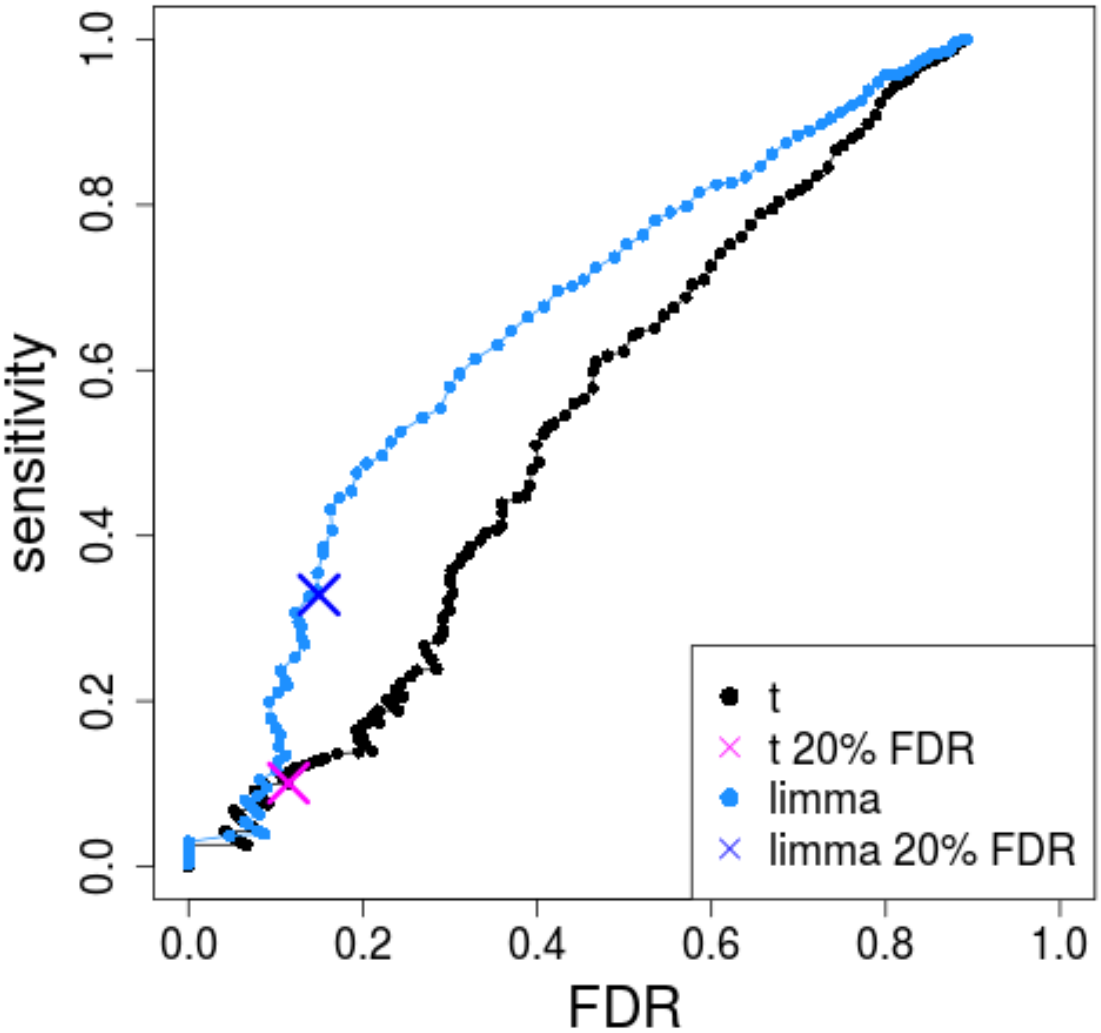


Figure 20: plot of chunk roc4

Summary

- limma provides a hierarchical model for moderation of variance estimates in the context of linear models
- Avoids false positives from under-estimation of variance
- Also addresses the gain in degrees of freedom from moderation

RNA-seq: counting molecules

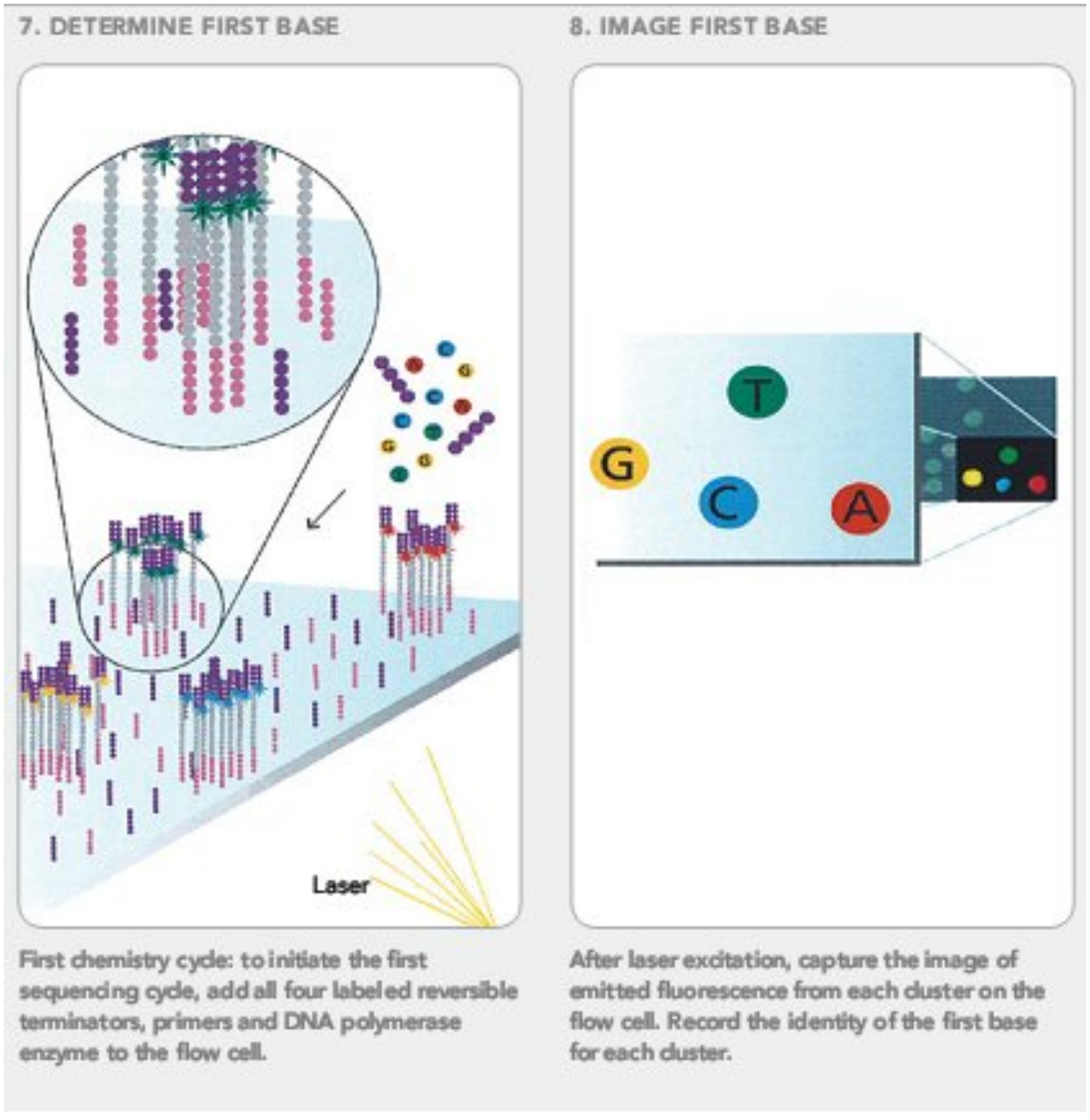


Figure 21:

RNA-seq: counting molecules

```
@SRR1265495.1 1/1
CTTTGCCCGCGTGTCAGACTCCATCCCTCCTCTGCCGCCACCGCAGCAGCCACAGGCAC
+
CCCFFFFFHGFHHHIJJIJIJJIJJJJJJJGIIJJJIJJJJJJJJJEFHDEFFFEECDDI
@SRR1265495.2 2/1
CCTGGCTGTGTCCATGTCAGAGCAATGGCCCAAGTCTGGGTCTGGGGGGGAAGGTGTCA
+
@C@FFFDFHHHHGGIIAGHI9GIIIIIGIIIIIIGI@@FHGDDDH@GGIBB05?B>ACCC
@SRR1265495.3 3/1
CTGTGTCCATGTCAGAGCAATGGCCCAAGTCTGGGTCTGGGGGGGAAGGTGTCATGGAC
+
@C@DFFFFGGHDHIIEGDHCGGHIJGEHIIJIIJIEEHGIGEGDDB@@@CDDDDEDDDI
```

for ~30 million reads (often pairs of reads)

RNA-seq: counting molecules

Align to genome or transcriptome

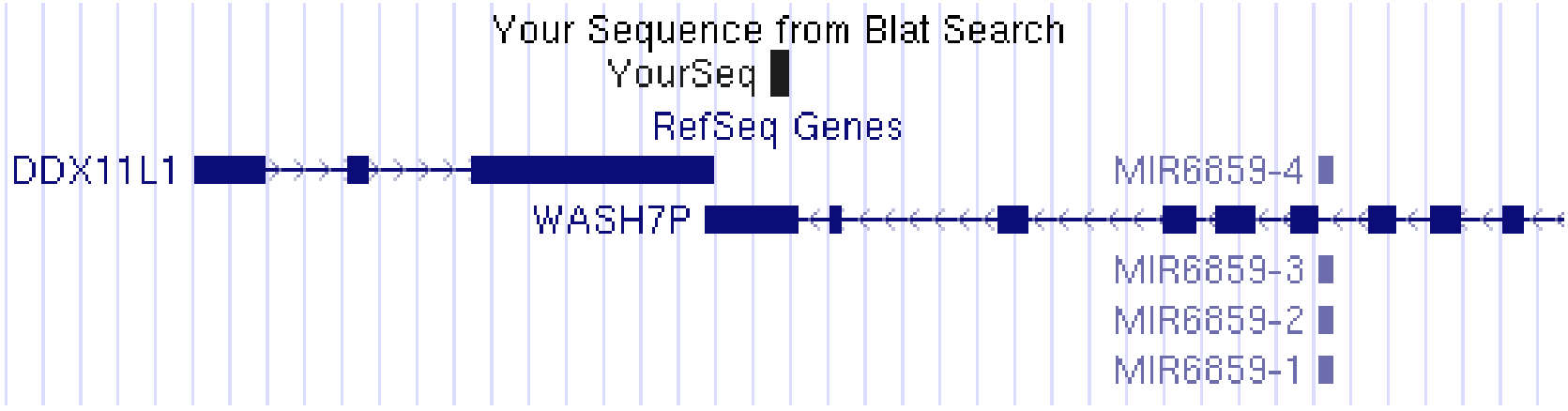


Figure 22:

RNA-seq: counting molecules

- Now for each gene and each sample we can obtain a *count* or *estimated count* of fragments
- Why estimated? Because some fragments cannot be uniquely associated with genes or isoforms
- Fast algorithms for probabilistically assigning:
- Salmon
- Sailfish (2014)
- kallisto (2016)
- Assume we have integer counts K_{ij} of unique fragments or from rounding estimated counts

Counts

1. Either model data with count distributions and inference with GLM

- DESeq2 (2014)
- edgeR (2010)
- many more

2. Learn weights associated with log normalized counts and use limma

- limma-voom (2014)

Most important for statistical analysis

- Total number of fragments is technical artifact
- Heteroskedasticity of counts
- Each gene has different variability

Total number of fragments

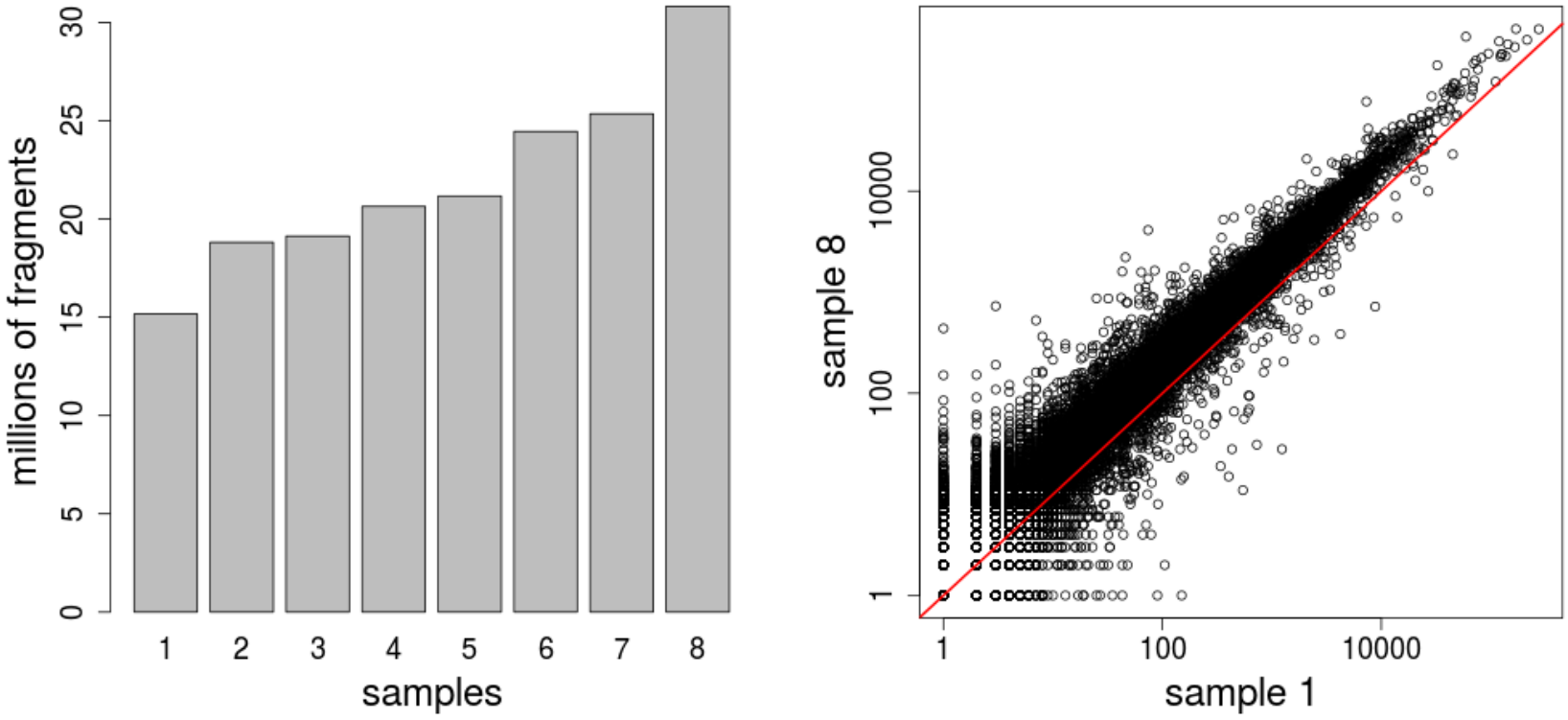


Figure 23: plot of chunk totalnumber

Sampling fragments

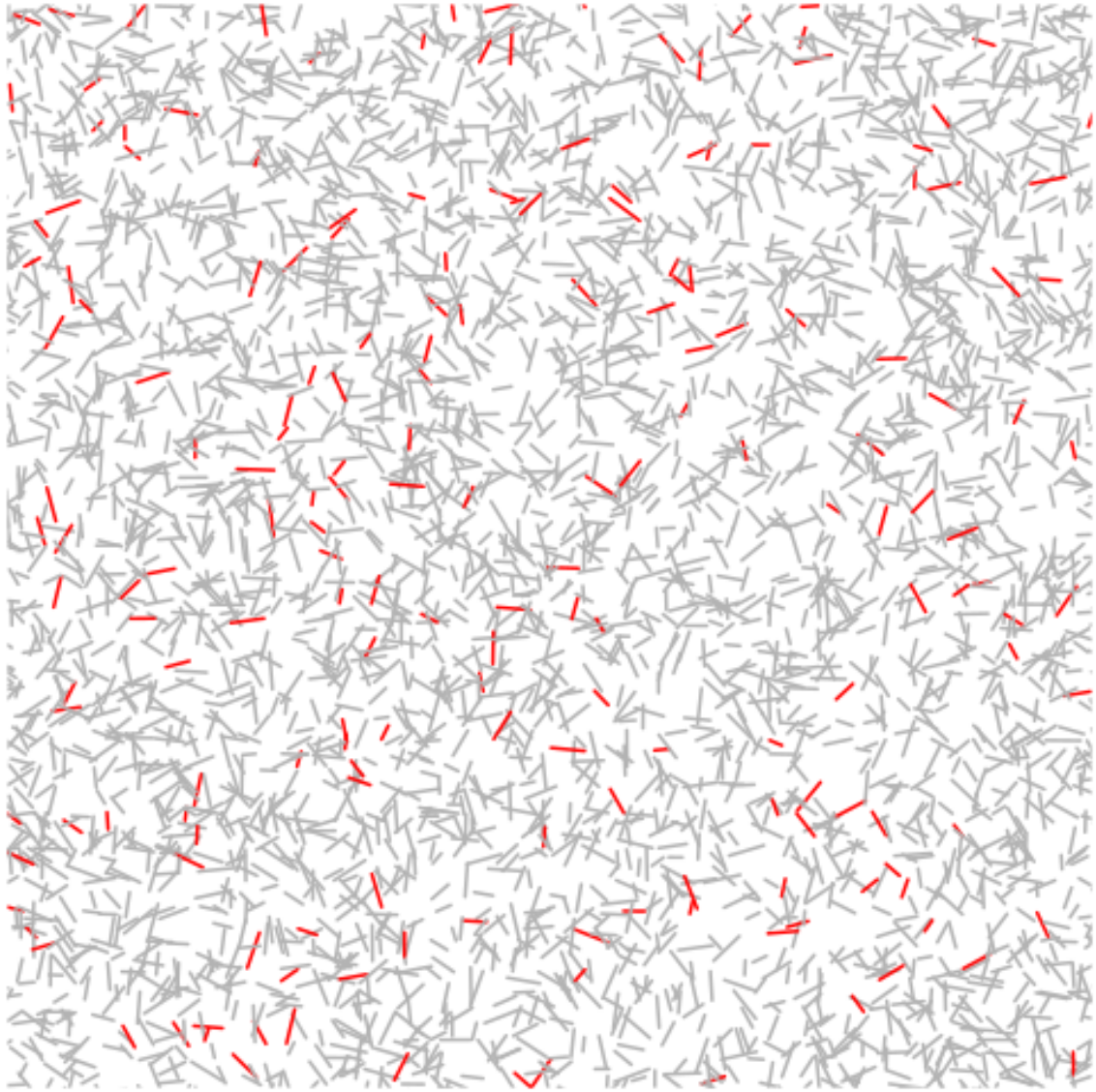


Figure 24: plot of chunk sampfrags

Poisson across technical replicates

- From Bullard 2010 take 7 technical replicates
- Calculate expected value $\hat{\lambda}_{ij}$ using DESeq2 norm
- $P(K_{ij} < \hat{\lambda}_{ij})$ assuming $K_{ij} \sim \mathrm{Pois}(\hat{\lambda}_{ij})$

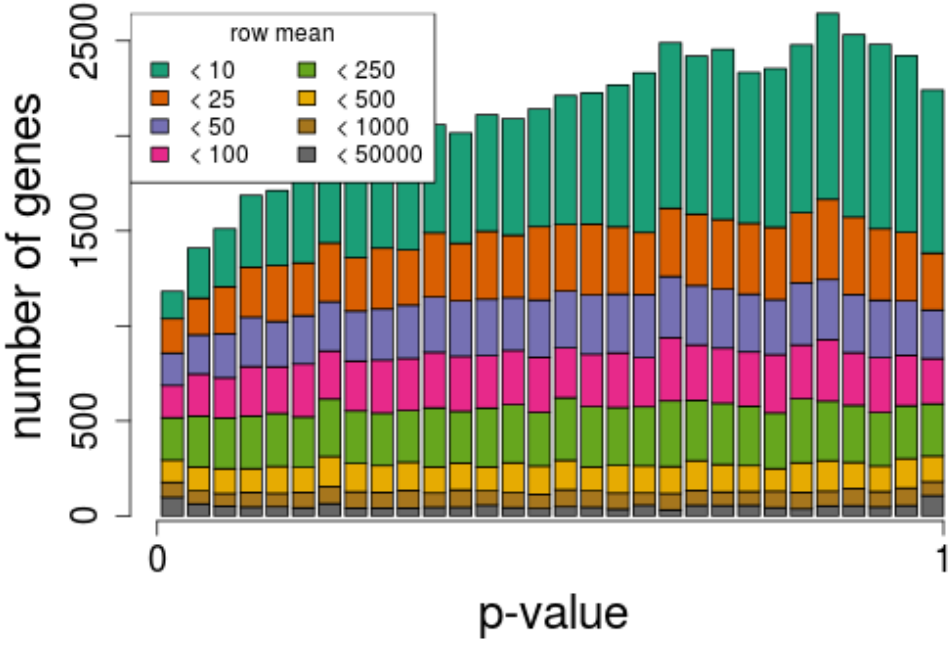


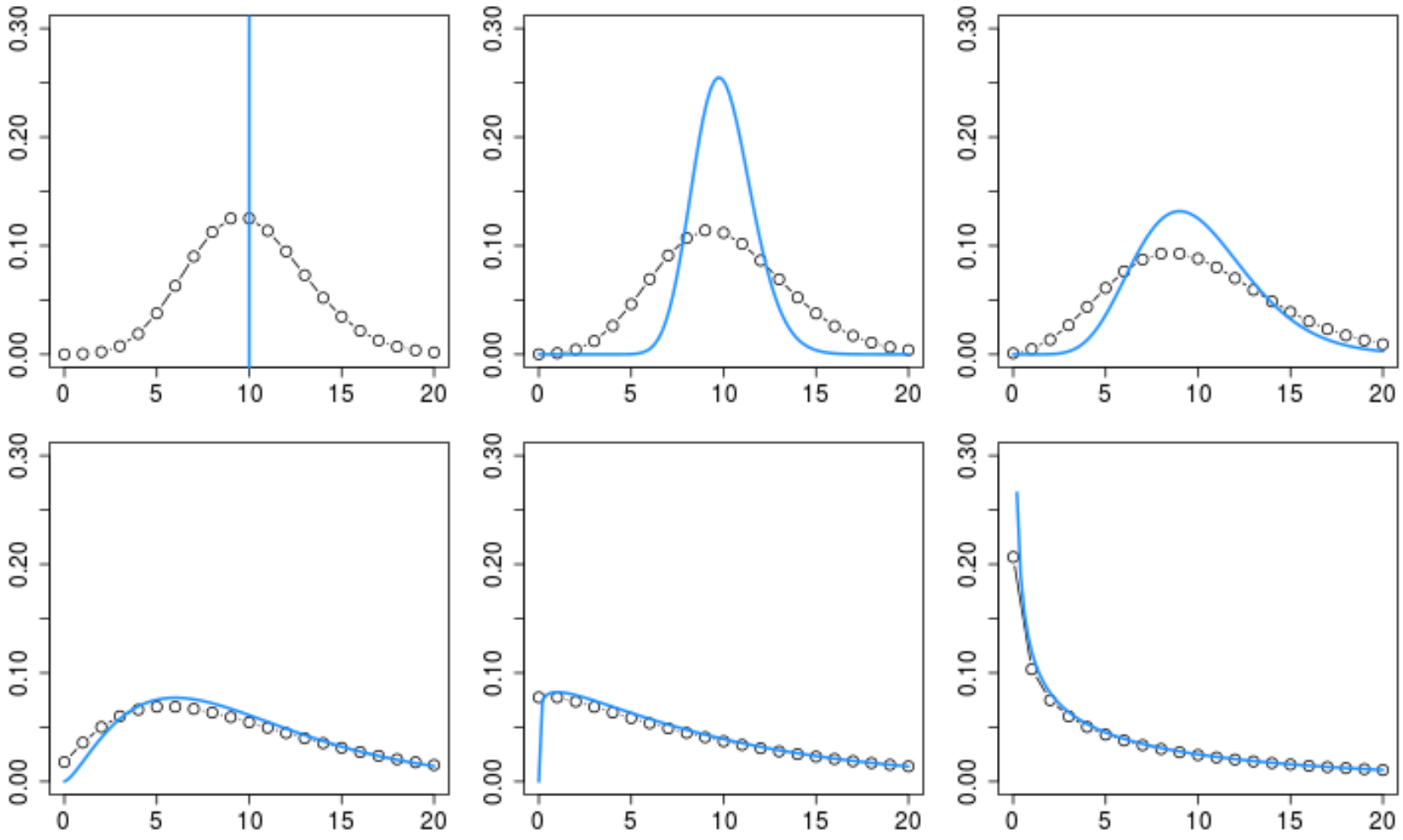
Figure 25: plot of chunk poisson

However, expression not equal across biological replicates

Figure 26:

Negative Binomial / Gamma Poisson

$$K_{ij} \sim \mathrm{NB}(\mu_{ij}, \alpha_i)$$
$$\mathrm{Var}(K_{ij}) = \mu_{ij} + \alpha_i \mu_{ij}^2$$



NB model for RNA-seq

- Similar to our microarray model for X_{ij}
- Added an s_j to deal with sequencing depth

$$
\begin{aligned}
K_{ij} &\sim \mathrm{NB}(\mu_{ij}, \alpha_i) \\
\mu_{ij} &= s_j q_{ij} \\
\log_2(q_{ij}) &= \beta_{i0}, \quad j \in A \\
\log_2(q_{ij}) &= \beta_{i0} + \delta_i, \quad j \in B
\end{aligned}
$$

$\delta_i \neq 0$ implies DE (differential expression)

Moderation of dispersion

- In DESeq2, a prior on $\log(\alpha_i)$
- Calculate the mean of normalized counts $\bar{\mu}_i$
- A trend line of dispersion over mean $\alpha_{tr}(\mu)$
- Width of the prior σ^2 estimated via assumption of normal sampling variance of $\log(\hat{\alpha}_i)$

$$\log(\alpha_i) \sim N(\log(\alpha_{tr}(\bar{\mu}_i)), \sigma^2)$$

Moderation of dispersion

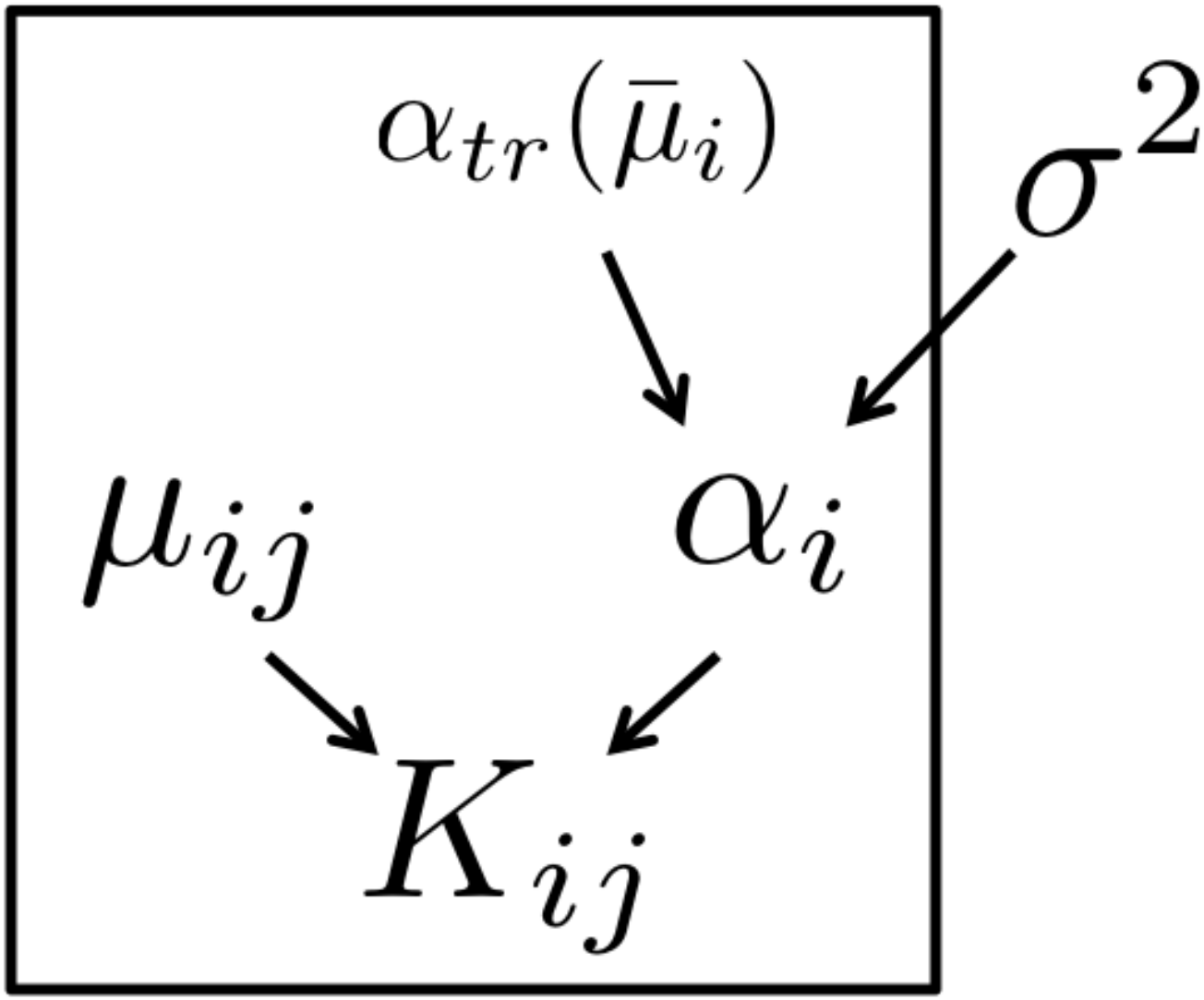


Figure 28:

Moderation of dispersion

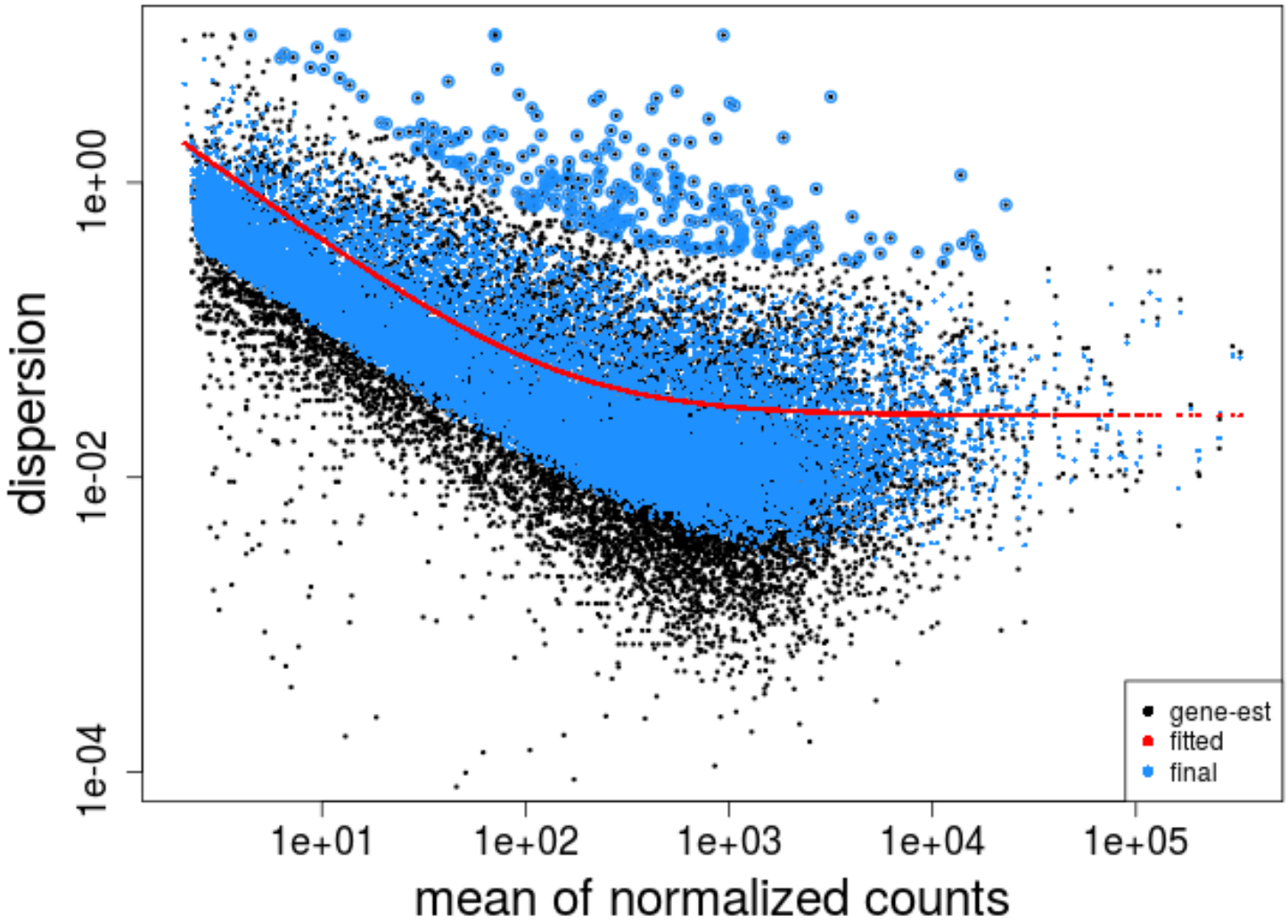


Figure 29: plot of chunk disp

Evaluate via simulation

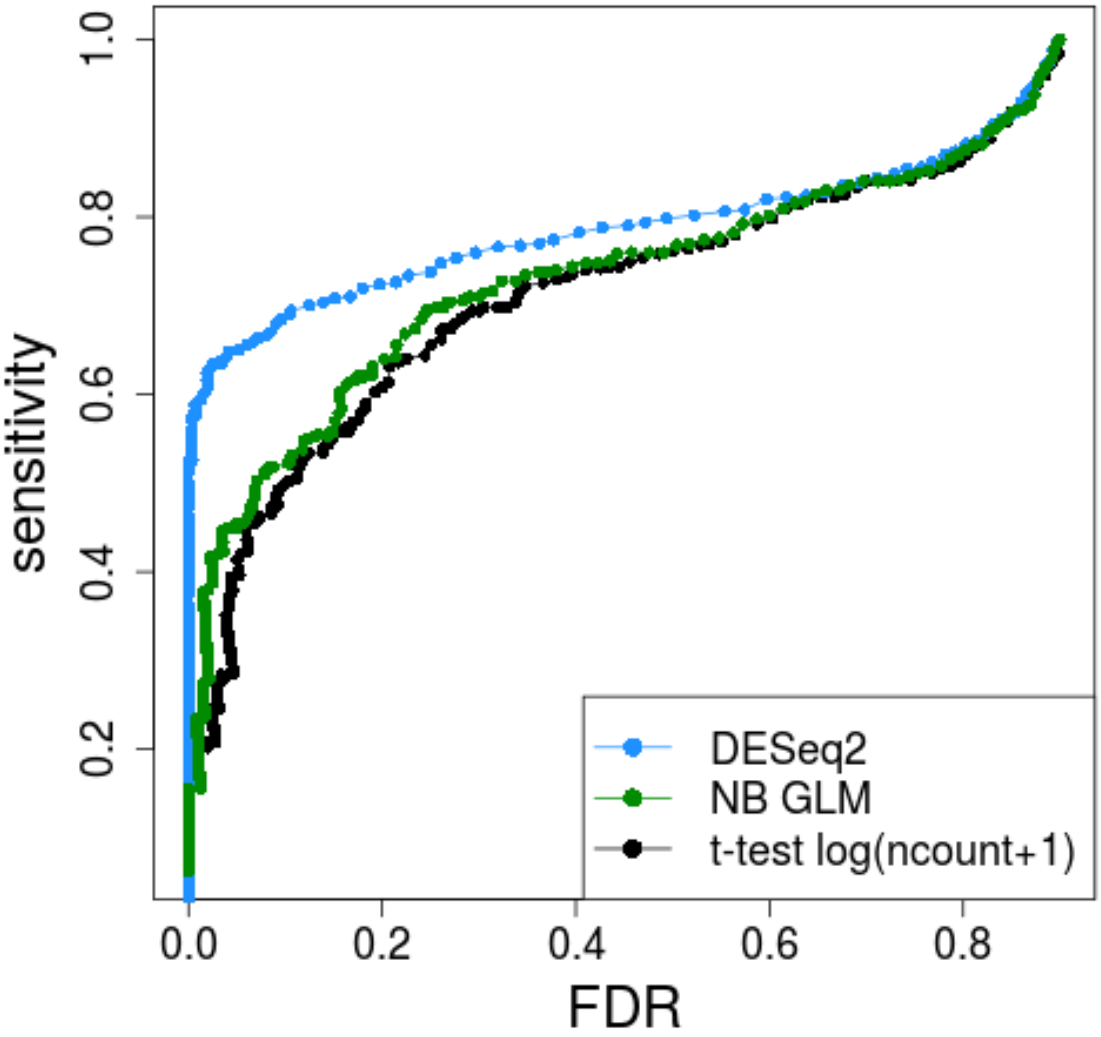


Figure 30: plot of chunk rnasim

Summary

- Model for counts similar to the hierarchical linear model, but constructed for the dispersion parameter
- Final dispersion estimate plug-in value in DESeq2, edgeR
- edgeR quasi-likelihood takes into account dispersion estimation uncertainty
- limma-voom uses weights on log normalized counts